Treatment Comparison with Survival and Non-survival Primary Endpoints

by

Shuo Xu

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

May, 2014

Abstract

Multiple biomarkers (surrogate endpoints) are often used to predict the failure event, as well as to evaluate the effect of treatment. In biomarker researches, Behrens-Fisher problem is a nonparametric two sample comparison problem, which is vital due to its generality. In this thesis, we proposed a new testing method for Behrens-Fisher problem. Our method dealt with multiple primary endpoints and focused on the global effect of a treatment rather than effect for every single endpoints, which separated us from most of the current studies. We reviewed existing methods dealing with multiple endpoints Behrens-Fisher Problem with complete data, and introduced basic characteristics and testing methods for survival data. In light of the limited ability to process censoring data with current methods, we created this new method combining information of non-survival markers and survival time to improve accuracy of the evaluation. Inspired by the idea of rank sum test or U statistics, we built an adjusted U-statistic to perform the hypothesis test for general two sample comparison problems. This test offered a reasonable approach to evaluate treatment effect and inform clinical decision-making when the length of follow-up is available and the

ABSTRACT

importance of the primary endpoints could be of equal-weight. In addition, the result of simulation indicated that our new test would have satisfactory performance under different real-world scenarios.

Primary Reader: Dr. Mei-Cheng Wang

Secondary Reader: Dr. Peng Huang

Acknowledgments

I would like to thank Dr. Mei-Cheng Wang, Dr. Peng Huang, and Dr. Chiung-Yu Huang for helpful discussion for this paper. Especially I will thank Dr. Mei-Cheng Wang for all the advice she has given me during the two years. Thank all my professors, they really taught me a lot. And thank my friends, especially Yue Chu, for always helping me and surpporting me in the hard time.

Dedication

This thesis is dedicated to my advisor, Dr. Mei-Cheng Wang, and to the place I have been for two years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Survival analysis is a critical component in biostatistics and biomedical research. But in many circumstances, the process of observing the failure events takes very long time, so we sometimes prefer to use one or more biomarker(s), which can predict the real endpoint, as the endpoint(s) instead of the real one, called surrogate endpoint(s). As surrogate endpoints, the biomarkers must reflect the real endpoint well. There are many researches in this area, generally Fleming and DeMets[1] have introduced some rules of choosing surrogate endpoints, such as, when we consider if a new treatment improves the surviving, the biomarkers we choose must be in the same biological pathway of how this treatment changes the surviving. As an example, in AIDS research, we usually use CD4 count as the surrogate endpoint instead of the real endpoint "developing AIDS". Fleming and DeMets[1] also mentioned several failures of trials because of bad surrogate endpoints.

Traditionally, the researchers used single surrogate endpoint. However, in many cases, single endpoint is not sufficient to address the sceintific question. As the requirements we need for valid surrogate endpoints, how to choose the endpoints became the most important question we are facing. Firstly, the true marker, maybe the first change in the body when a person is developing a disease, sometimes is not clinically observable or examinable. Such as many neuro diseases like Alzheimer's Disease, the change of amount reduction of neurons is not detectable before dissection. In this case, we can only use the measurable biomarkers as endpoints. And from this point of view, since any of them are not the true marker, or we say, there is not a single biomarker can predict 100 percent or near 100 percent sure of the failure event happening. For example, Huang et al.[2] suggested, Parkinson's disease disability was not able to be captured by a single outcome since it had many totally different aspects. Secondly, since there is not always a perfect marker for the failure as we stated before, if we would like to pick only one biomarker as the surrogate endpoints, there may be different biomarkers each has its advantages and disadvantages. Then there comes the realistic problem of consensus among investigators, different people may have different ideas and there is no right or wrong. Finally, since any single biomarker cannot capture the whole picture, we need more than one biomarkers to ensure the robustness. From all above, multiple endpoints are needed to capture the whole spectrum of prediction.

When we compare two treatment groups A and B, using multiple primary end-

points, it is often difficult to correctly assume the joint distributions for these endpoints, a better way is to do a nonparametric test. In most of the researches currently, the null hypothesis of the two sample test are

$$H_0 : F_1 = F_2$$

Where F_1 and F_2 are the underlying cumulative distribution functions of treatments A and B. That is, they assume the underlying distributions of the two groups are the same under the null hypothesis. But in the real world, the difference between different treatments will not be only a shift. In many cases, the biological pathways are even totally different between the two treatments, so we cannot assume the only thing changed is the mean value, there will probabily be many other changes that we do not care and will change the shape of the distribution. At this time, assuming that we only care about the mean value of the biomarker measurements rather than the whole shape, the null hypothesis above is not proper. This kind of problem is called Behrens-Fisher Problem[3], which means a general nonparametric problem of comparing two groups, without the assumption for the shapes of their distributions. For this problem, Brunner et al[4] raised a format of Hypothesis as below

$$H_0 : p = P(X < Y) + \frac{1}{2}P(X = Y) = \frac{1}{2}$$

where X, Y are the endpoint measurement random variables for subjects from groups A and B. Or equally, we can write $\theta = P(X > Y) - P(X < Y)$, then the null

hypothesis will be

$$H_0 : \theta = 0$$

This format offers a very good intuition and a big flexibility, and will fit more general questions[5, 6].

On the other hand, in the multiple endpoints problem, there is a general question must be asked, that is, what is the decision strategy, or how do we decide when the influence of the two treatments on different endpoints are different. Because it is almost impossible that one method beats the other one on each of the endpoints we measured, the decision is not obvious when the two treatments both have its better parts. In many cases, for example in Parkinson's disease's treatments, it has many different important aspects that we cannot say that there is one aspect that is important and another one is not. Under this condition, Huang et al.[2] raised a new format of hypothesis that put equal weight on all of K biomarkers, that is,

$$H_0 : \theta = \sum_{v=1}^{K} \theta_v / K \leq 0 \quad v.s. \quad H_1 : \theta > 0,$$

where $\theta_v = P(X_v > Y_v) - P(X_v < Y_v)$ and X_v, Y_v are the random variables of the vth biomarker measuments for subjects from groups A and B. This setting makes an easy and reasonable decision making strategy by looking at the global effect, when we have multiple endpoints with equal importance.

In most of trials focusing on biomarkers, as we mentioned before, people were dealing with biomarkers as surrogate endpoints or predictions for the true failure, instead of using the true failure. But there are many cases that we could also observe

some true failure. In this situation, it is very helpful if we could add the real failure time into consideration. On the other hand, "failure event" can be encountered in different scenarios For example the initiation of levodopa is a failure event in Parkinson disease progression. Motor Function scores can be masked by the use of levodopa. Since the different patients have different levodopa initiation times, the follow up time will be different if we compare motor function score. Therefore, if we just measure the biomarkers without considering the followup time, the results may be biased. So we would like to add the real failure as another endpoint among the biomarkers and see the whole picture. But when we are discussing survival data, we must deal with censoring. In this work, we will also discuss the methods dealing with Behrens-Fisher Problem with censored data, and try to combine all the endpoints together to be one test.

Chapter 2

Analytical Methods

2.1 Current Research with Complete Data

As we are considering Behrens-Fisher Problem, we do not care the shapes of the underlying distributions, we set our null hypothesis only about the mean. For convenience, we assume that for each outcome, the larger value is preferred[7]. If the original measurement is not like that, we will do transformations to ensure this assumption.

Let $X_i = (X_{i1}, ..., X_{ik})$ be the k outcomes from subject $i(i = 1, 2, ..., m)$ in group A, and $Y_i = (Y_{j1}, ..., Y_{jk})$ be the k outcomes from subject $j(j = 1, 2, ..., n)$ in group B. Suppose the X_i's are independent and identically distributed, with joint cumulative distribution function F, and the Y_j's are independent and identically distributed,

with joint cumulative distribution function G. The null hypothesis is set as below [2]:

$$H_0 : p_v = P(X_{iv} < Y_{jv}) + \frac{1}{2}P(X_{iv} = Y_{jv}) = \frac{1}{2}$$

where $v = 1, 2, ..., k$. Or equally, for the vth outcome let $\theta_v = P(X_{iv} > Y_{jv}) - P(X_{iv} < Y_{jv})$, the null hypothesis can be written as following [2]:

$$H_0 : \theta_1 ... = \theta_k = 0.$$

2.1.1 O'Brien's Test

Under nonparametric assumption, we usually use the rank or rank sum to get information and to test if two groups have the same mean. If we consider the multiple biomarkers case, assuming we have k biomarkers, we are able to get ranks for each of the biomarkers. For the vth outcome, we combine the two groups together and the $m + n$ subjects $X_{1v}, ..., X_{mv}, Y_{iv}, ..., Y_{nv}$, and then a unique rank will be assigned to each subject. Then the ranks are seperated into two groups with respect to the original groups, denoteed as $R_{x,iv} = rank(X_{iv}), R_{y,jv} = rank(Y_{jv})$, if we havs tied data, the rank of the tied subjucts will defined as the average of their ranks. The total rank for ith subject in group A is defined as $R_{xi} = \sum_{v=1}^{k} R_{x,iv}$, similarly for the jth subject in group B, the total rank is $R_{yj} = \sum_{v=1}^{k} R_{y,jv}$.

Consider equal variance case, O'Brien[8] performed a test as a regular two-sample t-test for two rank-sum samples: $R_{x1}, ..., R_{xm}$ and $R_{y1}, ..., R_{yn}$, with the standard

deviation estimated by pooled sample standard deviation. Denote

$$\bar{R}_x = \frac{\Sigma_{i=1}^m R_{xi}}{m}, \qquad \bar{R}_y = \frac{\Sigma_{j=1}^n R_{yj}}{n}$$

$$\hat{\sigma}_x^2 = \frac{1}{m-1}\Sigma_{i=1}^m (R_{xi} - \bar{R}_x)^2, \quad \hat{\sigma}_y^2 = \frac{1}{n-1}\Sigma_{j=1}^n (R_{yj} - \bar{R}_y)^2, \quad \hat{\sigma}^2 = \frac{(m-1)\hat{\sigma}_x^2 + (n-1)\hat{\sigma}_y^2}{m+n-2}.$$

The O'Brien's test statistic is defined as

$$T_1 = \frac{\bar{R}_y - \bar{R}_x}{\hat{\sigma}\sqrt{1/m + 1/n}}.$$

2.1.2 Adjusting of O'Brien's Test

The O'Brien's test is very widely used in clinical trails since it is simple in format and easy to calculate, but when we go back to our assumptions of the problem as a Behrens-Fisher Problem, this test actually has a big problem such as it is not always correct when people use it.

In the paper of Huang et al.[7], they proved that assuming $m/n \to \lambda$ as $N = m+n \to \infty$ for some finite constant λ, under the null hypothesis, as N goes to infinity, the O'Brien's test statistic T_2 converges in distribution to a normal distribution with mean zero and variance h_2, where h_2 are well defined as below:

$$h_1 = \frac{\sum_{u=1}^k \sum_{v=1}^k (1+\lambda)^2 (a_{uv} + b_{uv}\lambda)}{\sum_{u=1}^k \sum_{v=1}^k [e_{uv}\lambda^3 + (b_{uv} + 2f_{uv})\lambda^2 + (a_{uv} + 2q_{uv})\lambda + p_{uv}]}.$$

Where

$$a_{uv} = cov(G_u^o(X_u), G_v^o(X_v)),$$

$$b_{uv} = cov(F_u^o(Y_u), F_v^o(Y_v)),$$

$$e_{uv} = cov(F_u^o(X_u), F_v^o(X_v)),$$

$$f_{uv} = cov(F_u^o(X_u), G_v^o(X_v)),$$

$$p_{uv} = cov(G_u^o(Y_u), G_v^o(Y_v)),$$

$$q_{uv} = cov(G_u^o(Y_u), F_v^o(Y_v)).$$

and $F_u^o(t) = P(X_u < t) + \frac{1}{2}P(X_u = t), G_u^o(t) = P(Y_u < t) + \frac{1}{2}P(Y_u = t)$, are the mid-distribution function for the uth outcome $(u = 1, 2, ..., k)$.

Consider the underlying distributions of two groups, which are F and G, we notice that the null hypothesis does not imply $F = G$. When the null hypothesis is true but F and G are different, the h_1 will not equal to 1, which ruins the statement that the O'Brien's test statistic T_1 follows the t distribution. Therefore, the type I error will not be controlled if we use O'Brien's test when we have unequal underlying distribution.

Under this circumstance, Huang et al.[7] has proposed a method to adjust the variance term in the O'Brien's test statistic to make them follow t distribution, as following,

$$T_{1a} = \frac{\bar{R}_y - \bar{R}_x}{\hat{\sigma}\sqrt{\hat{h}_1(1/m + 1/n)}}.$$

And the simulations by Huang et al.[7] showed a good performance of the adjusted statistic.

2.2 Introductions of Characteristics of Survival Data

Survival analysis is one of the major branches in biostatistics, and it is unique and very important due to the special characteristics of survival time random variables. It deals with the analysis of the time to a failure event, called failure time or survival time. This failure event is the event we set as an endpoint, such as death, diagnosis of disease, or hospitalization, it can be either repeatable or not. In survival analysis, the interest of research is usually the surival function and hazard function. The survival function is defined by,

$$S(t) = P(T \geq t),$$

where T is the survival time from the time origin to the failure event, it is a non-negative random variable, and t is any timepoint. When T is a continuous random variable, we have $S(t) = 1 - F(t)$. And hazard function is defined by

$$\begin{aligned} \lambda(t) &= P(T = t | T \geq t) \quad (if\ \ T\ \ is\ \ discrete), \\ &= \lim_{\Delta \to 0^+} \frac{P(t \leq T < t + \Delta | T \geq t)}{\Delta} \quad (if\ \ T\ \ is\ \ continuous), \end{aligned}$$

which has the property of $\lambda(t) = f(t)/S(t)$. When T is continuous, the harcard function is the instantaneous failure rate at t given survival until t. Here $f(t)$ is the probability density function of T and $F(t)$ is the cumulative density function of T. In many cases, for example when T follows exponential distribution, the hazard function

and the survival function will have very simple expressions which make good sence for interpretation and for real world cases.

The uniqueness of surival analysis is that it comes with unique formats of missing data. The most common format of missing data is called right-censoring, which means we lose followup of some subject at a time point. Censoring can due to many reasones, such as subjuect's withdrawal, death due to other reasons, or the end of the study (administrative censoring). In most cases, we assume the censoring time C is independent of failure time T.

In real study, the observed time is $X = min(T, C)$, paired with binary censoring indicator δ. The censoring indicator equals to 1 when the subject is not censored, eqauls to 0 otherwise. Because of censoring, we cannot estimate the survival function of the failure time with the methods dealing with complete data. A widely used estimator for survival function is Kaplan-Meier Estimator[10], and it will deduce to empirical distribution when there is no censoring. For two-sample testing, we usually perform log-rank test and Gehan's test. In our work, Gehan's approach will be applied on our test statistics, and detailed discussion will be presented in the following section.

2.3 Test With both Usual Biomarkers and Survival Time

Since it is very hard to estimate the joint distribution of different biomarkers and survival time. Instead of estimating the joint distribution directly, we focus on the distribution of their ranks alternatively. There are a lot of studies dealing with multiple biomarkers, but almost none of them includes survival time information. Here we use U-statistics, a kind of unbiased statistic, to combine all the information together. All the biomarkers and survival time are coded as the larger value is preferred. Also we assume that independent censoring in our case.

Using the similar notions with in section 2.1. Assume we have K biomarkers and we also care about the survival time. Let $X_i = (X_{i1}, ..., X_{iK}, X_{i,K+1}, \delta_{Ai})$ be the observation from subject i (i=1,2,..,m) in group A, and $Y_j = (Y_{j1}, ..., Y_{jK}, Y_{j,K+1}, \delta_{Bj})$be the observation from subject j (i=1,2,..,n) in group B. Here X_{iv} or Y_{jv}, $v = 1, 2, ..., K$ is the measurements for one subject of the vth biomarker. $X_{i,K+1}$ or $Y_{j,K+1}$ is the survival time observation with $X_{i,K+1} = min(T_{Ai}, C_{Ai})$ and $Y_{j,K+1} = min(T_{Bj}, C_{Bj})$, where T is for true failure time and C is for censoring; and δ_{Ai}, δ_{Bj} is the censoring indicator. Denote G_1, G_2 are the "survival function" of the censoring time C_1, C_2.

We are testing the hypothesis as below:

$$H_0 : \theta = \sum_{v=1}^{K+1} \theta_v/(K+1) = 0$$

$$v.s.$$

$$H_1 : \theta \neq 0,$$

where for the vth endpoint, $\theta_v = P(X_{iv} > Y_{jv}) - P(X_{iv} < Y_{jv}), \quad v = 1, 2, ..., K + 1.$

Consider the vth non-survival outcome, we rank all the subjects in both groups together and assume the rank sum of group A is R_{Av}, by the results of Mann-Whitney, R_{Av} can be written as

$$R_{Av} = \frac{m(m+n+1)}{2} + \frac{1}{2}U_v$$

where $U_v = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij,v}$, and

$$U_{ij,v} = 1 \qquad if \quad X_{iv} > Y_{jv},$$

$$U_{ij,v} = 0 \qquad if \quad X_{iv} = Y_{jv},$$

$$U_{ij,v} = -1 \qquad if \quad X_{iv} < Y_{jv}.$$

Or we can write,

$$U_{ij,v} = I(X_{iv} > Y_{jv}) - I(X_{iv} < Y_{jv})$$

Under the null hypothesis, U_v will follow a Normal distribution with mean 0 and variance calculable. Here, for each variable, the difference between the rank sum of group A and the corresponding U-statistic is a constant as the group sizes for both group are fixed. Therefore, we could build the statistics using the U-statistics, instead of the original rank sum, and get the same result. On the other point of view, the

definition of this statistic give us a good "description" on how to estimate the θ in the original question, where

$$\theta = \sum_{v=1}^{K+1} \theta_v/(K+1)$$

And

$$\theta_v = P(X_{iv} > Y_{jv}) - P(X_{iv} < Y_{jv}).$$

Similarly, for the survival outcome, we would like to use the same statistic to build a link to the rank-sum of a specific group or the value of θ which we care the most. So here, for combining the multiple endpoints, we would like to add all the rank sums together. Also we adjust the statistic with the sample size to make it a good estimation of the θ, we will calculate the "adjusted total U sum"

$$U_{sum} = \left(\sum_{v=1}^{K} U_v + U_{K+1}\right) \Big/ ((K+1)mn) \quad .$$

As all these U-statistics are following normal distributions, so the adjusted sum U_{sum} will still follow the normal distribution with mean 0 and calculable standard error.

For survival time variable, it is hard to gain the ranks of real survival time of T_{Ai}'s and T_{Bj}'s because of censoring, as we mentioned earlier in this work. So we need to estimate the U_{K+1}. There are two approaches we can use here, the first one is to estimate the value of the U-statistic with the same way as in Gehan's Test. That is,

$$\hat{U}_{K+1} = \sum_{i=1}^{m}\sum_{j=1}^{n} \hat{U}_{ij,v},$$

where

$$\hat{U}_{ij,K+1} = 1 \qquad if \quad T_{Ai} > T_{Bj},$$

$$\hat{U}_{ij,K+1} = 0 \qquad if \quad T_{Ai} = T_{Bj}, or \ \ unknown$$

$$\hat{U}_{ij,K+1} = -1 \qquad if \quad T_{Ai} < T_{Bj} \qquad .$$

i.e.

$$\hat{U}_{ij,K+1} = I(T_{Ai} > T_{Bj}) - I(T_{Ai} < T_{Bj})$$

For the alternative approach, we use the method introduced in Cheng, Wei, and Ying's paper[9] to estimated $I(T_{Ai} > T_{Bj})$, that is, estimating $I(T_{Ai} > T_{Bj})$ by

$$\frac{\delta_{Bj} I(X_{i,K+1} > Y_{j,K+1})}{\hat{G}_1(Y_{j,K+1})\hat{G}_2(Y_{j,K+1})}.$$

Where $\hat{G}_1$ and $\hat{G}_2$ is the estimated survival function for censoring, using Kaplan-Meier Estimator. And we use the same method to estimate $I(T_{Ai} < T_{Bj})$. Then obviously we can also estimate the corresponding statistic U_{K+1} with this estimation. That is,

$$\hat{U}_{ij,K+1} = \frac{\delta_{Bj} I(X_{i,K+1} > Y_{j,K+1})}{\hat{G}_1(Y_{j,K+1})\hat{G}_2(Y_{j,K+1})} - \frac{\delta_{Ai} I(X_{i,K+1} < Y_{j,K+1})}{\hat{G}_1(X_{i,K+1})\hat{G}_2(X_{i,K+1})} \quad .$$

We can prove that this statistic is a U-statistic by the definition of two-sample U-statistics. So,

$$\hat{U}_{K+1} = \sum_{i=1}^{m}\sum_{j=1}^{n} \hat{U}_{ij,K+1} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(\frac{\delta_{Bj} I(X_{i,K+1} > Y_{j,K+1})}{\hat{G}_1(Y_{j,K+1})\hat{G}_2(Y_{j,K+1})} - \frac{\delta_{Ai} I(X_{i,K+1} < Y_{j,K+1})}{\hat{G}_1(X_{i,K+1})\hat{G}_2(X_{i,K+1})} \right) \quad .$$

Until now, we have obtained all the U-Statistics for biomakers and estimated the U-statistic for the survival time variable using two different approaches. we can calculate the estimated "adjusted total U sum" for all variables for group A, as

$$\hat{U}_{sum} = \left(\sum_{v=1}^{K} U_v + \hat{U}_{K+1} \right) \Big/ ((K+1)mn) \quad .$$

Under the null-hypothesis, the theoretical mean of $\hat{U}_{sum}$ is 0, and we will use bootstrap to estimate the variance in each specific case when we have real data.

Further more, we still would like to consider the theoretical value of the variance of the total U sum $\hat{U}_{sum}$. Let

$$W_v = N^{1/2} \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij,v}/(2mn),\ v = 1, ..., K$$

$$W_{K+1} = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\omega(\beta) \frac{\delta_{Bj} I(X_{i,K+1} > Y_{j,K+1})}{\hat{G}_1(Y_{j,K+1}) \hat{G}_2(Y_{j,K+1})} - \omega(-\beta) \frac{\delta_{Ai} I(X_{i,K+1} < Y_{j,K+1})}{\hat{G}_1(X_{i,K+1}) \hat{G}_2(X_{i,K+1})} \right).$$

Here $\omega(.)$ is a weight function, when we set it $\omega(.) = 1$ we get $W_{K+1} = \hat{U}_{K+1}$. In paper of Huang et al.[4], they proved, with some trivial assumptions, under the null-hypothesis, the vector $(W_1, ..., W_K)$ converges to a multivariate normal distribution with mean $(0, ..., 0)$ and variance matrix which has the following form:

$$\Sigma = \frac{N(n_2 - 1)}{n_1 n_2} A + \frac{N(n_1 - 1)}{n_1 n_2} B + \frac{N}{4 n_1 n_2} C,$$

where

$$a_{uv} = Cov(F_{2u}(X_{.u}), F_{2v}(X_{.v}))$$

$$b_{uv} = Cov(F_{1u}(Y_{.u}), F_{2v}(Y_{.v}))$$

$$c_{uv} = Cov(\xi(X_{.u}, Y_{.u}), \xi(X_{.v}, Y_{.v})),$$

and here, she assumed that $\{(X_{i1}, ..., X_{iK}), i = 1, ..., m\}$ and $\{(Y_{j1}, ..., Y_{jK}), j = 1, ..., n\}$ are independent identically distributed realizations of the random vectors $(X_{.1}, ..., X_{.K})$ and $(Y_{.1}, ..., Y_{.K})$; $F_{1u}(t) = P(X.u < t) + P(X.u = t)$, $F_{2v}(t) = P(Y.v <$

$t) + P(Y.v = t)$; and $\xi(x, y) = I(x > y) - I(x < y)$. And the estimation method for this covariance matrix was also mentioned by Huang et al.[4].

Chapter 3

Simulations and Examples

3.1 Simulations

To evaluate the performance of the test we described above, we simulate the original data from different distributions. Assume both group have 100 individuals, and there are K biomarkers, $K = 3$. Also, assume that the length of the study period, which is the maximum possible value of censoring time, is 3 years. For the population we generated, we use Normal distribution and Beta to generate the biomarkers data, exponential distribution to generate the true survival time. Also, we consider the correlation between the covariates, as well as generate the simplest case that all the covariates are independent. For censoring time, we generate it from an independent uniform distribution along the study period, regarding the independent censoring assumption.

Each time after generating a set of data, we use both of the two approaches we mentioned in the last section to obtain the statistic $\hat{U}_{sum}$'s on the same data. We use bootstrap to estimate the standard deviation and 95% confidence interval. At the same time, for each distributions we use to generate data, we will generate 500 sets of data, with each we calculate the variance and confidence interval of the statistics $\hat{U}_{sum}$'s, to explore or compare the properties of the variance by the two methods, also for the cases under the null hypothesis H_0 we check the Type I errors and for cases under alternative hypothesis H_1 we check the powers of the test.

Here we consider 3 pairs of settings of the true underlying distribution of the endpoints and we evaluate the performance of the two approaches of test under each of the settings. In each pair, we have one settings under the null hypothesis and one under the alternative hypothesis. And the two settings in the same pair has the same type of distrubution for each of the biomarkers, but they will have different mean vector. In the three cases under althernative hypothesis, we also perform a Bonferroni test (denoted by Approach 3 in the summary table) to compare the power and type I error of the tests. Here are the 3 pair of cases, denoted by case 1-3,

Case 1 Biomarkers data from groups A and B follows multivariate normal distribution, with mean μ_1, μ_2 and shared covariance matrix

$$\Sigma_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Survival time for groups A and B are generated from exponential distributions $Exp(\lambda_1)$, $Exp(\lambda_2)$. The censoring time for each subject from each group are following the Uniform distribution on the interval $[0, 3]$. For the under null hypothesis setting, $\mu_1 = \mu_2 = (0, 0, 0)$, $\lambda_1 = \lambda_2 = 0.5$. For the under althernative hypothesis setting, $\mu_1 = (0, 1, 2)$, $\mu_2 = (0, 1.2, 2.5)$, $\lambda_1 = 0.5$, $\lambda_2 = 0.4$.

Case 2: In groups A and B, biomarkers data follows multivariate normal distribution, with mean μ_1, μ_2 and the same covariance matrix

$$\Sigma_3 = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix}.$$

The survival time for each subject from each groups A and B are generated from exponentials distribution $Exp(\lambda_1)$, $Exp(\lambda_2)$ plus 0.03 times of the value of measurement of the second biomarker. That is,

$$T_{Ai} = T'_A + 0.03X_{i2}, \quad T'_A \sim exp(\lambda_1)$$

$$T_{Bj} = T'_B + 0.03Y_{j2}, \quad T'_B \sim exp(\lambda_2)$$

The censoring time for both group are following the Uniform distribution on the interval $[0, 3]$. For the under null hypothesis setting, $\mu_1 = \mu_2 = (0, 0, 0)$, $\lambda_1 = \lambda_2 = 0.5$. For the under althernative hypothesis setting, $\mu_1 = (0, 1, 2)$, $\mu_2 = (-1, 3, 2.5)$, $\lambda_1 = 0.5$, $\lambda_2 = 0.4$.

Case 3: In group A, biomarkers follow multivariate normal distribution, with mean

μ_1 and covariance matrix Σ_3, where Σ_3 is same defined as in case 2. In group B, the data of 3 biomarkers are following Beta distributions $Beta(0.7, 1)$, $Beta(0.5, 1)$, $Beta(2, 4)$, correspondingly, plus a shared error ϵ, where ϵ follows normal distribution with mean 0 and standard deviation 0.05. That is,

$$Y_{j1} = Y'_{j1} + \epsilon, \quad Y'_{j1} \sim Beta(0.7, 1)$$

$$Y_{j2} = Y'_{j2} + \epsilon, \quad Y'_{j2} \sim Beta(0.5, 1)$$

$$Y_{j3} = Y'_{j3} + \epsilon, \quad Y'_{j3} \sim Beta(2, 4)$$

$$\epsilon \sim N(0, 0.05^2)$$

. Survival time for each subject from groups A and B are generated, from from exponential distributions $Exp(\lambda_1)$, $Exp(\lambda_2)$ plus 0.03 times of the value of measurement of the first biomarker. The censoring time for both group are following the Uniform distribution on the interval $[0, 3]$. For the under null hypothesis setting, $\mu_1 = (0.41, 0.33, 0.33)$, according to the mean of Beta distributions in group B, $\lambda_1 = \lambda_2 = 0.5$. For the under althernative hypothesis setting, $\mu_1 = (-0.5, 0, 0)$, $\lambda_1 = 0.5$, $\lambda_2 = 0.4$.

In case 1, 2, we run the bonfferoni test by Wilcoxon's rank sum test for biomarkers and Gehan's test with bootstrap with survival time. In case 3, since the underlying distributions of biomarkers in two groups have different shapes, it will be biased if we still use Wilcoxon's rank sum test without adjusting, then here we use bootstrap

methods to perform the Bonfferoni test. The simulation results are shown in the table 3.1.

From the simulation results, we can see both the two approaches perform well in different kinds of underlying distribution settings. On average, the first approach has a slightly smaller variance than the second approach. They both offer a good contorlled Type I error around 0.05, especially Approach 2. For the power term, Approach 2 has a slightly lower power than Approach 1, which may be improved if we adjust for the Kaplan-Meier estimator before we calculate the test statistic in Approach 2. Compared with Bonfferoni test, Although the Bonfferoni test sometimes has a increditable high power, we need to notice that what we are testing is the global effects. Observing the case settings, it is easy to see that two groups may have big difference for one sigle biomarker, but the data is coded as that "difference between the groups" may have different directions, or significant difference for only one biomarker will not be that significant averagely for the whole picture. In this situation, the high power of Bonfferoni test will not be meaningful. On the other hand,our new approaches make more sense in the setting of the global wised null hypothesis. Because, for a beferroni test, the null hypothesis is usually "each biomarkers has same mean bewteen the two groups", instead of our global setting. Therefore, our new apporach is better fitted in the goobal testing problem. And this results indicated that the performance of the test have not been influenced by the shape of the underlying distribution, that means when we are not able to specify the same kind of distribution or same shape between

Table 3.1: simulation summary.

Hypothesis	Case	Approach	Mean($\hat{U}_{sum}$)	SD($\hat{U}_{sum}$)	Type I error	Power
H_0	1	1	0.00137	0.0378	0.038	
		2	0.00157	0.0437	0.038	
		3			0.038	
H_1		1	-0.0962	0.0375		72.2%
		2	-0.0967	0.0427		63.8%
		3				81.6%
H_0	2	1	0.000851	0.0442	0.060	
		2	0.000974	0.0496	0.046	
		3			0.054	
H_1		1	-0.212	0.0371		68%
		2	-0.212	0.0421		61.6%
		3				100%
H_0	3	1	-0.00114	0.0485	0.054	
		2	-0.00158	0.0532	0.052	
		3			0.050	
H_1		1	-0.179	0.0467		96.6%
		2	-0.179	0.0510		94.3%
		3				98.6%

the variables in the two groups, it is a very good test to justify which treatment has the high expectation value.

In the above simulation, we assumed that the distributions of the censoring time are the same between the two groups. When the censoring distributions are different between two groups, we can not use the approach 1 because of the model assumption reqiurement for Gehan's Method of estimating the U statistic. But this unequal censoring case is also a general problem in the real studies. So for the same distribution types on biomarkers and survival times as in case 3, we just change the censoring distribution of group B to be exponential distribution $Exp(0.5)$ bounded by $[0, 3]$, to gain the two new settings denoted as **case 4**. In case 4, for the under null hypothesis setting, $\mu_1 = (0.41, 0.33, 0.33)$, according to the mean of Beta distributions in group B, $\lambda_1 = \lambda_2 = 0.5$. For the under althernative hypothesis setting, $\mu_1 = (0, 0, 0)$, $\lambda_1 = 0.5$, $\lambda_2 = 0.4$. In case 4, Under this two setting we run the simulation to check the perform of the second approach and we got table 3.2. Notice that here we use bootstrap method for biomarkers and the same mathod with Approach 2 for survival times when we run the Bonfferoni test.

Comparing table 3.2 with table 3.1, we find that this test also performs well in the unequal censoring conditions. Despite that the Bonfferoni test will produce incorrect high power, when for each single biomarker the difference is samll but there is a significant global difference, as the setting in case 4, the power of Bonfferoni test also will go low, but in this situation our new approach has sensitively detected the global

Table 3.2: simulation summary for unequal censoring cases.

Hypothesis	Case	Approach	Mean($\hat{U}_{sum}$)	SD($\hat{U}_{sum}$)	Type I error	Power
H_0	4	2	-0.00749	0.0531	0.060	
H_0		3			0.054	
H_1		2	-0.129	0.0519		68%
H_1		3				53.6%

difference. But when we run it for many times we will find that sometimes the type I error will be be controlled as well as the same censoring distribution cases. This may be caused to the unadjustment of the estimated survival function for censoring time for each distribution. In this test, we just plug in the Kaplan-Meier estimator without adjusting, which may cause bias of estimating U_{K+1} in certain cases. For example, the value of Kaplan-Meier estimator may be very small on the tails of the curve, which will cause large variance when we calculate the $\hat{U}_{K+1}$. In the future, we will do the adjustment for Kaplan-Meier estimator to improve the performance of our testing method.

3.2 Example

To mimic the real data, we generate three sets of data from case 3, 4 mentioned in the simulation section, with sample size equals to 200 in each group and numbers of endpoints $K + 1 = 4$. Treat them as the original data we have, we perform the

test based in these data.

3.2.1 Example 1

Firstly we generate a set of data with setting case 3 under null hypothesis, sample size $m = n = 200$ and number of endpoints $K + 1 = 4$. We first use histograms to explore the basic characteristics of the data we have for both groups (Figure 3.1 and Figure 3.2). From the histograms we could see that the distrbution shapes of each biomarkers are different from the two groups, but they will have the about the same mean.

Then we perform the test by both of the approaches, and Table 3.3 shows the results we got from the two approach. From here we can see, the results from the two approaches are very similar and the first approach has a slightly smaller variance and slightly smaller difference with the theoretical mean. But on the whole the two tests performs both very well here.

Table 3.3: Example 1: Testing results.

Approach	$\hat{U}_{sum}$	$SD(\hat{U}_{sum})$	95% C.I.for $\hat{U}_{sum}$	Decision
Approach 1	-0.0260	0.0345	[-0.0959, 0.0337]	Fail to reject
Approach 2	-0.0325	0.0378	[-0.110, 0.0382]	Fail to reject

Also we plot the histogram of $\hat{U}_{sum}$'s (Figure 3.3) by the two approaches and we can see it estimated the value of θ well.

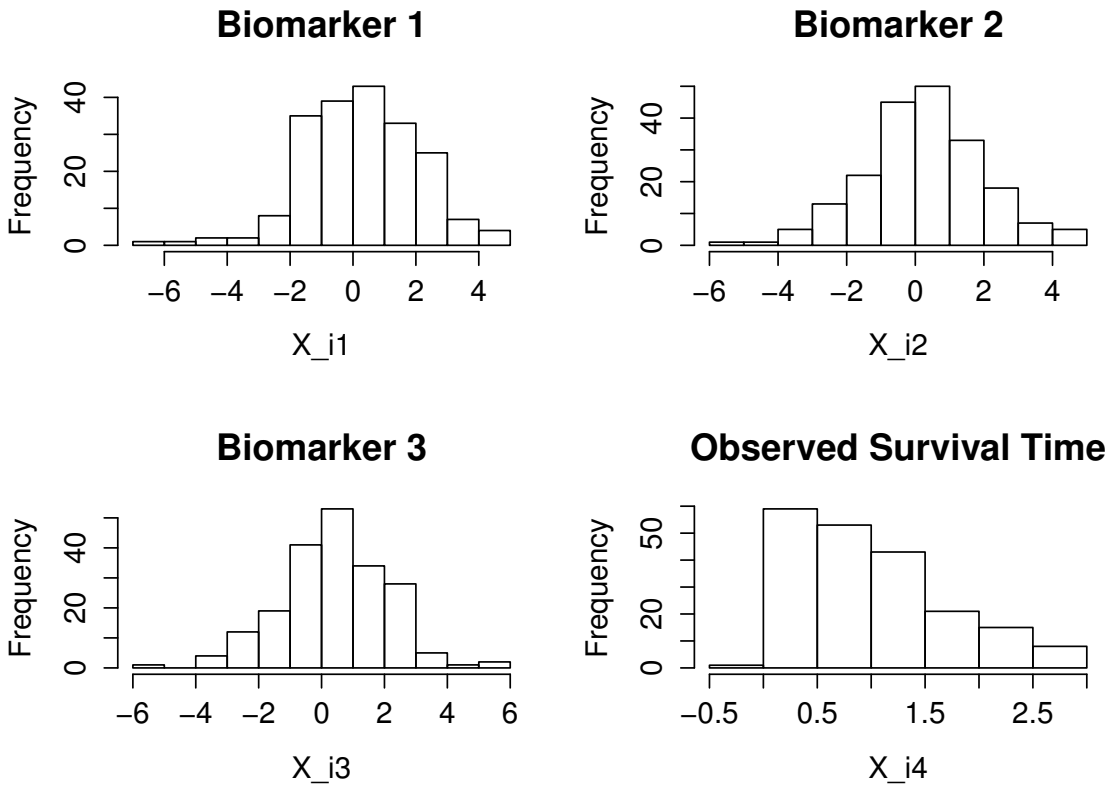

Figure 3.1: Example 1: data distribution of group A (histograms).

3.2.2 Example 2

For the second example we generate a set of data with setting case 3 with alternative hypothesis, sample size $m = n = 200$, $\mu_1 = (0, 0, 0)$ and and number of endpoints $K + 1 = 4$. Similarly, we use histograms (Figure 3.4 and Figure 3.5) to explore the basic characteristics of the data we have for both groups. Notice that the underlying distribution is under the alternative hypothesis.

We perform the test by both of the approaches, the summary of results we got

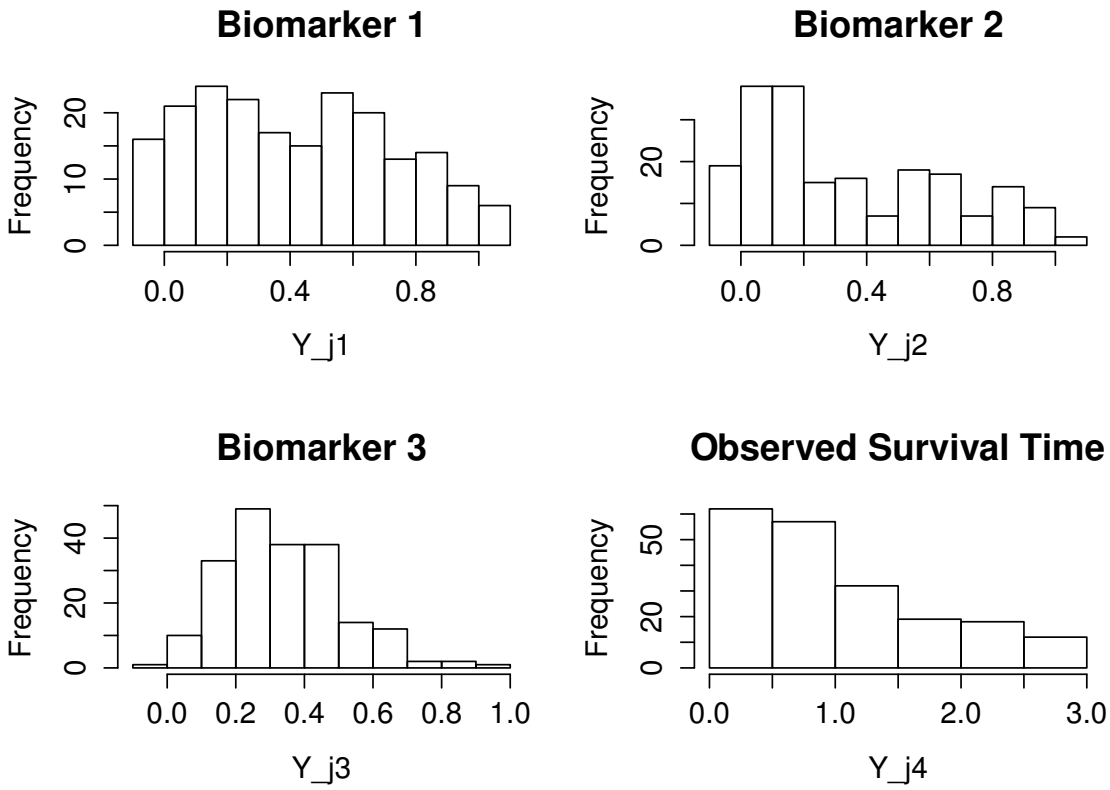

Figure 3.2: Example 1: data distribution of group B (histograms).

from the two approach is listed in Table 3.4. At the same time we have the histogram of $\hat{U}_{sum}$ (Figure 3.6). From here we can see, the results from the two approaches are still very similar and the first approach has a slightly smaller variance. Both the two tests performs both very well here. And from Figure 3.6, we can find the $\hat{U}_{sum}$'s from both of approaches has estimated the value of θ well.

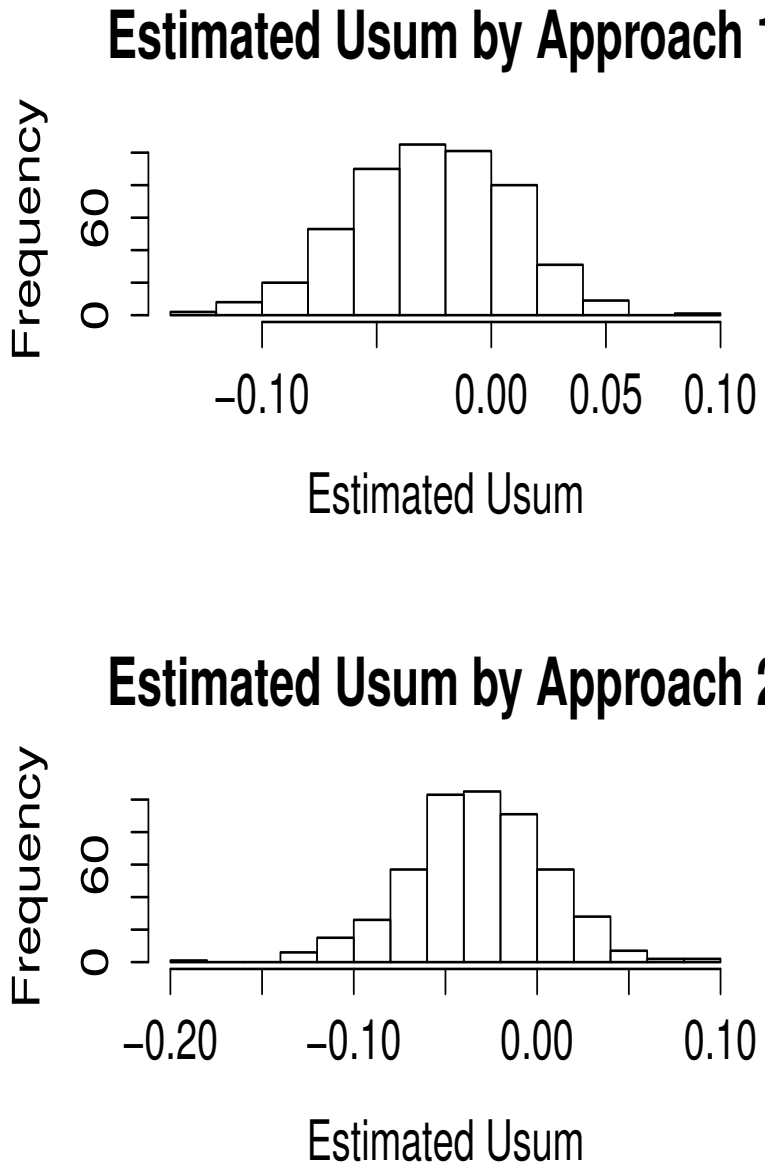

Figure 3.3: Example 1: Distribution of $\hat{U}_{sum}$, Approach 1,2.

3.2.3 Example 3

Finally we generate a set of data with setting case 4 with alternative hypothesis, sample size $m = n = 200$ and number of endpoints $K + 1 = 4$. The histograms of the

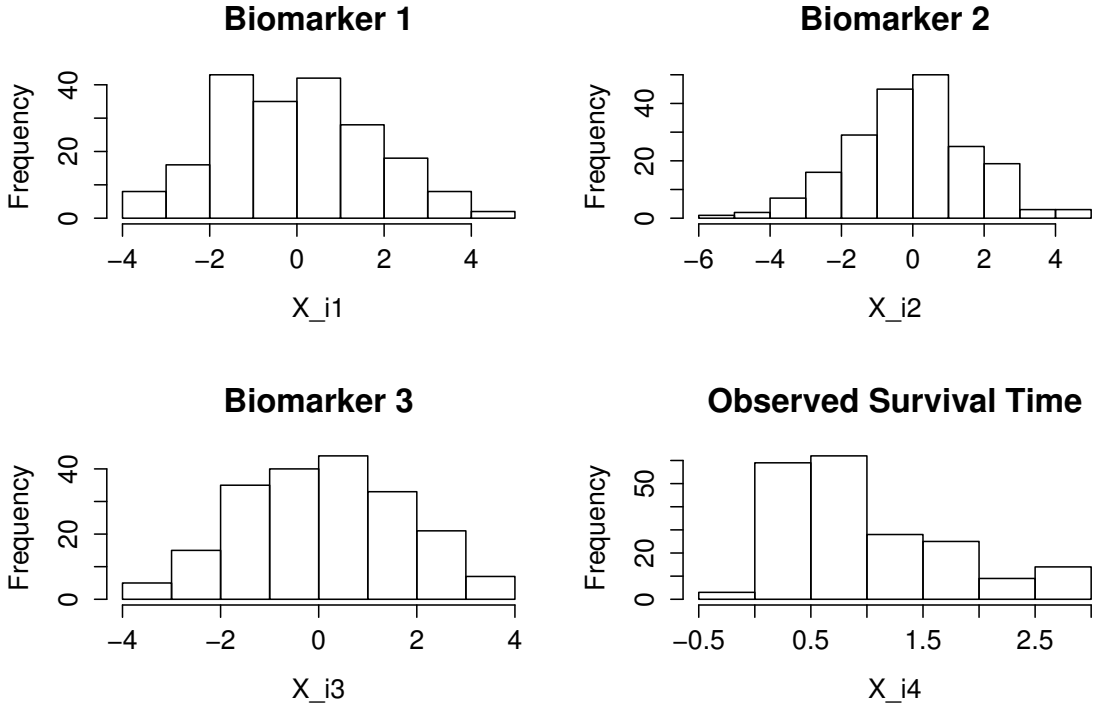

Figure 3.4: Example 2: data distrubution of group A (histogram).

Table 3.4: Example 2: Testing results.

Approach	$\hat{U}_{sum}$	SD($\hat{U}_{sum}$)	95% C.I.for $\hat{U}_{sum}$	Decision
Approach 1	-0.138	0.0327	[-0.202, -0.0756]	Reject
Approach 2	-0.144	0.0376	[-0.218, -0.0631]	Reject

four endpoints for both groups is shown in Figure 3.7 and Figure 3.8. Notice that the in this setting alternative hypothesis is true.

As we mentioned before, in this unequal consering condition, we can only use the second approach. We list the reault of the test on this data in Table 3.5 and histogram

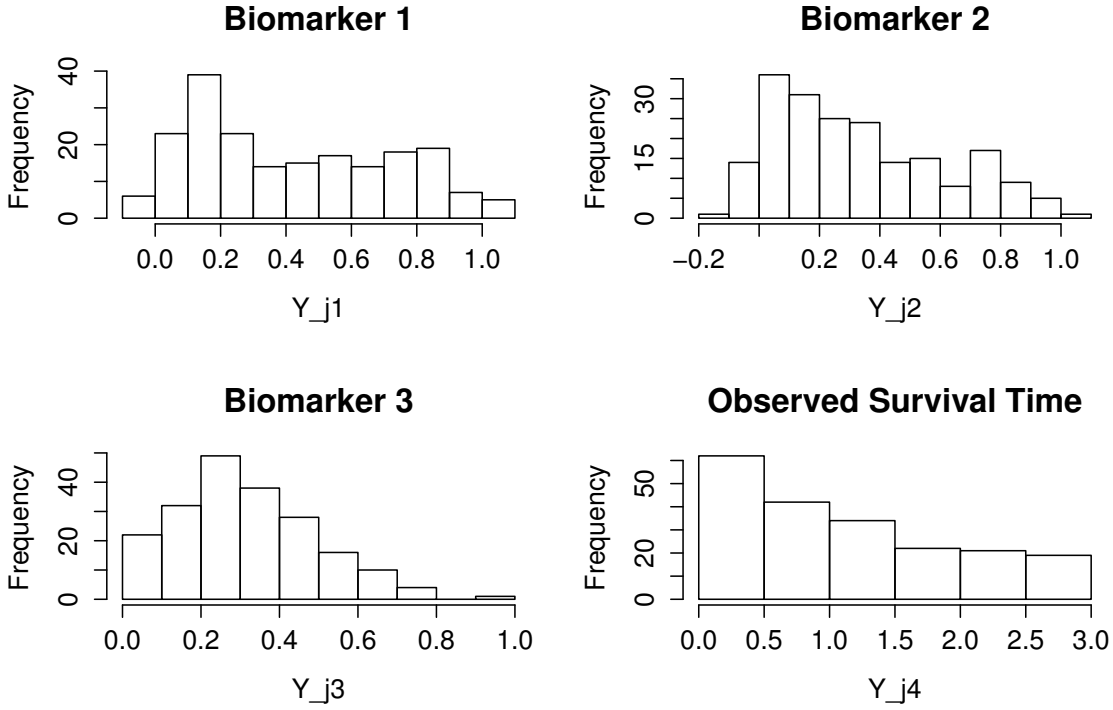

Figure 3.5: Example 2: data distrubution of group B histogram).

of $\hat{U}_{sum}$ in Figure 3.9.

Table 3.5: Example 3: Testing results.

Approach	$\hat{U}$	SD($\hat{U}_{sum}$)	95% C.I.for $\hat{U}_{sum}$	Decision
Approach 2	-0.144	0.0347	[-0.212, -0.0673]	Reject

From the plotting (Figure 3.9) we could see the estimation of U_{sum}'s by each time of the bootstrap process are well normally distributed, and by the result summary we could see this method works well in this example case and the estimated variance term of the estimated U, or we say the estimation of θ is fairly well.

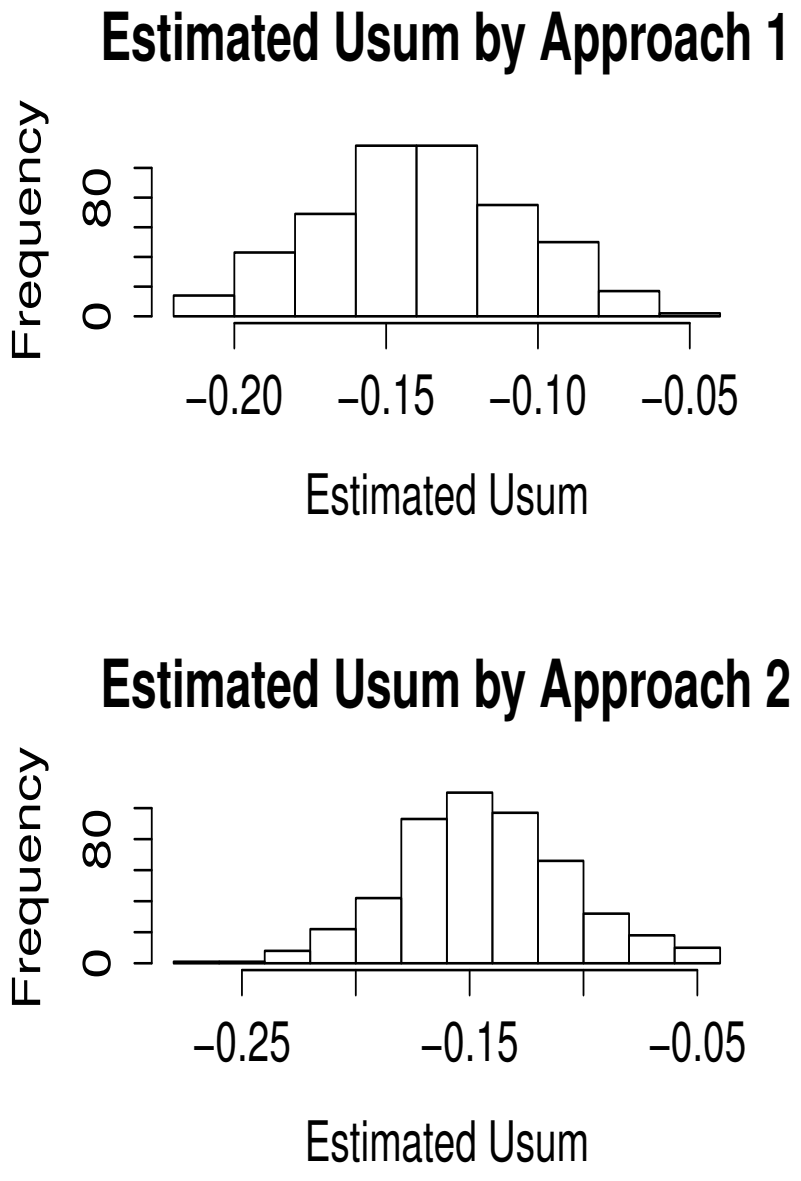

Figure 3.6: Example 2: Distribution of $\hat{U}_{sum}$, Approach 1,2

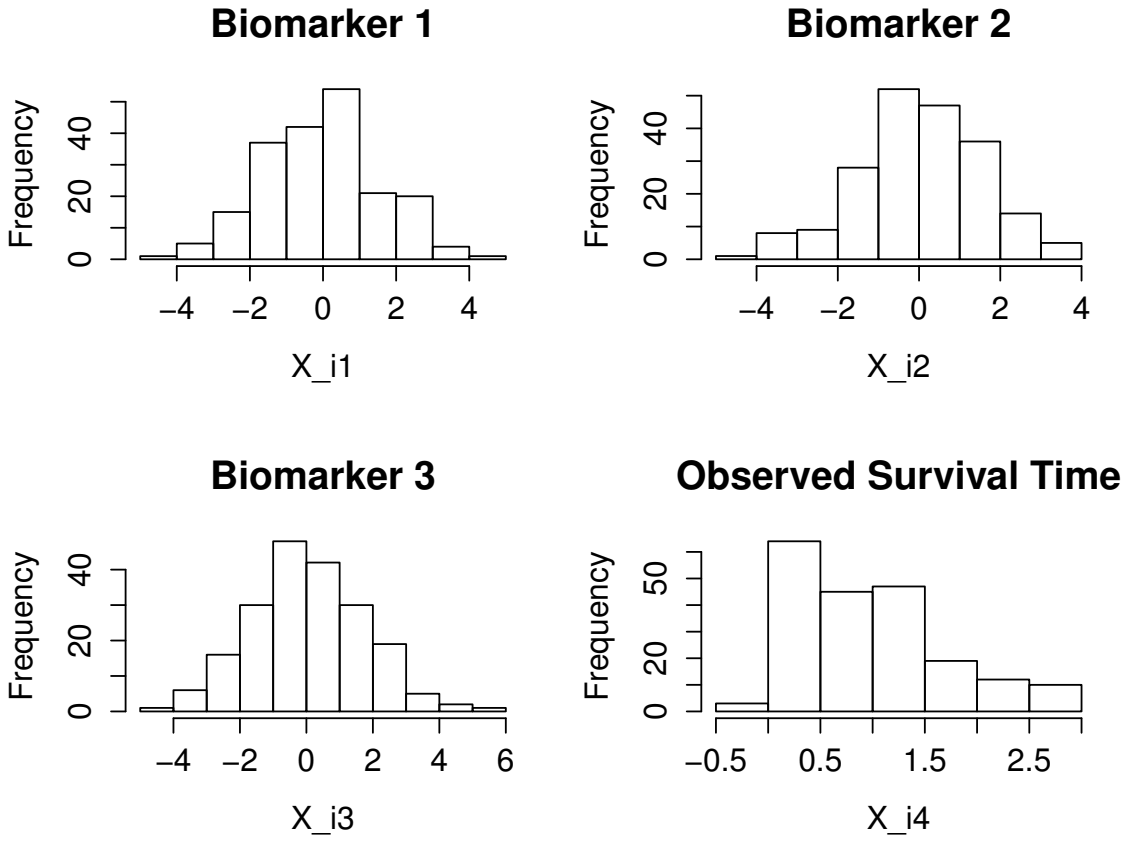

Figure 3.7: Example 3: data distrubution of group A (histogram).

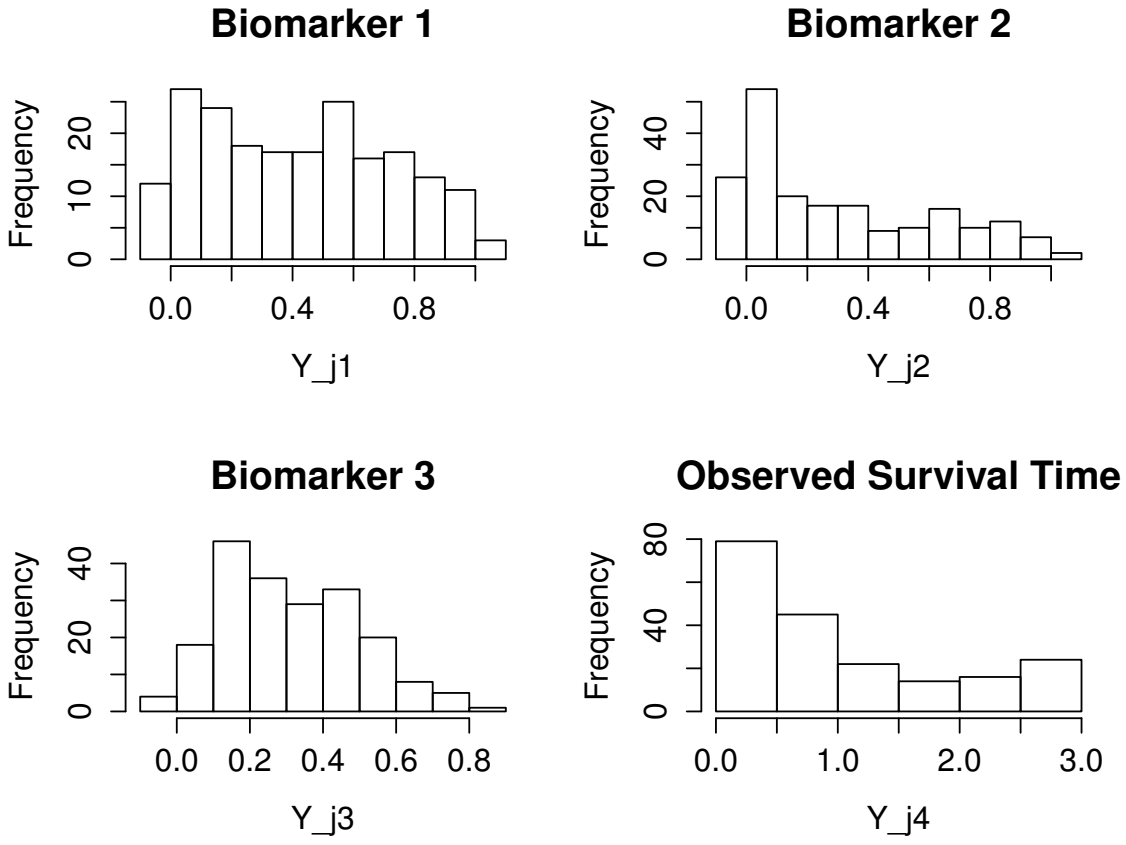

Figure 3.8: Example 3: data distrubution of group B (histogram).

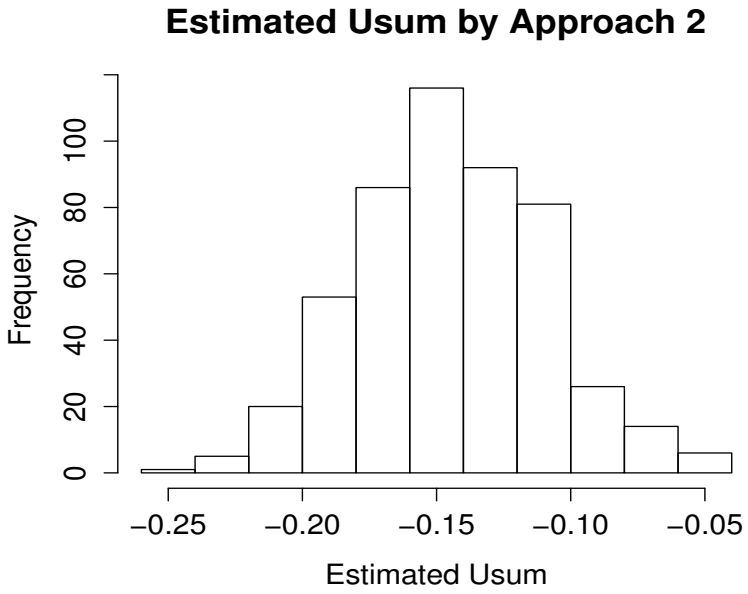

Figure 3.9: Example 3: Distribution of $\hat{U}_{sum}$, Approach 2.

Chapter 4

Discussion

The rank sum test for multiple endpoints showed a very practical way to deal with Behrens-Fisher Problem and is already widely used in clinical trials research. Now we put survival time into the test methods, offers a good method to make better use of the survival information we have and will definitely increase the accuracy of the comparison.

The big advantage of this method is that, as we mentioned in both method and simulation sections, it is non-parametric and can deal with the two sample problems without assuming the same underlying assumptions, which will fit more general problems. Also, different from many traditional tests, we do not assume that between the two groups the directions of changes of each biomarker are all the same, which means we do not require one group beating or being not worse than the other on every aspects before we say this treatment is better. Instead, this method offers a way to

evaluate a treatment globally on a whole picture.

From the simulation section, we could see that in different conditions or assumptions, we can get the correct decisions by performing this test using bootstrap at most of time. From both approaches of test, we can ensure a high power and low Type-I-error. But, the correlation we assumed in the simulation part is very simple, the real case will be more complicated. We still need to put more effort on how to mimic the correlation in the real data, as the correlation between biomarkers and between biomarker and survival time has a lot of different formats.

For the first approach, we use the Gehan's Method of estimating the U statistics, which requires the equal censoring distribution for both groups. This gives this method a big limitation when we try to use it in a real study. In many cases, we will have informative censoring, or the censoring time may depend on different not medical related condition, which may lead to a difference between the censoring distribution of the two groups. From this viewpoint, the second approach will be more general and requires less assumptions.

At the same time, the developing of the theoretical expression of our new statistic and its property is going to be completed in the close future. Also, as mentioned in simulation section, work still need to be done to adjusting for the Kaplan-Meier estimator for approach 2, will may be able to offer a better performance of this approach.

At this point, we see advantages and feasibility of this new testing method. This

method offers a good way to solve the problem of the lack of analysis for survival time in biomarkers research and a way to make more sense in treatment comparison in many real situations by evaluating global effects . We have reasons to believe that our new approaches will be useful and have a good performance on real data, and we believe the test will be developed more completely in the future.

Appendix A

R code

A.1 Bootstrap Function for Approach 1 (with Bonferroni Test for Survival Time)

```
Bootstrap1<-function(data10,data20,X10,X20,ind10,ind20,n1,n2,N,K){

  U<-rep(0,500)
  Us<-rep(0,500)

  for (J in 0:499)
  {
    num1<-c(1:n1)
    num2<-c(1:n2)
    sam1<-sample(num1, size=n1, replace = TRUE, prob = NULL)
    sam2<-sample(num2, size=n2, replace = TRUE, prob = NULL)

    x<-data10[sam1,]
    y<-data20[sam2,]
    array1<-array(rep(x,n2),dim=c(n1,K,n2))
    arrayx<-aperm(array1,c(1,3,2))
    array2<-array(rep(y,n1),dim=c(n2,K,n1))
```

```
    arrayy<-aperm(array2,c(3,1,2))

    Ubio<-(arrayx>arrayy)-(arrayx<arrayy)

    xs<-X10[sam1]
    ys<-X20[sam2]
    xi<-ind10[sam1]
    yi<-ind20[sam2]
    matx<-matrix(rep(xs,n2),ncol=n2)
    matix<-matrix(rep(xi,n2),ncol=n2)
    maty<-matrix(rep(ys,n1),nrow=n1,byrow=T)
    matiy<-matrix(rep(yi,n1),nrow=n1,byrow=T)

    Usur<-(matix&matiy)*((matx>maty)-(matx<maty))-
    (matix&(!matiy))*(matx<=maty)+((!matix)&matiy)*(matx>=maty)

    Us[J+1]<-sum(Usur)/(n1*n2)
    U[J+1]<-(sum(Ubio)+sum(Usur))/((K+1)*n1*n2)
  }

  return(c(mean(U),sd(U),quantile(U, probs = c(0.025,0.975),na.rm=T),
  quantile(Us, probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)))

}
```

A.2 Bootstrap Function for Approach 1 (with Bonferroni Test)

```
Bootstrap1<-function(data10,data20,X10,X20,ind10,ind20,n1,n2,N,K){

  U<-rep(0,500)
  UB<-matrix(NA,nrow=500,ncol=4)

  for (J in 0:499)
  {
    num1<-c(1:n1)
    num2<-c(1:n2)
```

```
    sam1<-sample(num1, size=n1, replace = TRUE, prob = NULL)
    sam2<-sample(num2, size=n2, replace = TRUE, prob = NULL)

    x<-data10[sam1,]
    y<-data20[sam2,]
    array1<-array(rep(x,n2),dim=c(n1,K,n2))
    arrayx<-aperm(array1,c(1,3,2))
    array2<-array(rep(y,n1),dim=c(n2,K,n1))
    arrayy<-aperm(array2,c(3,1,2))

    Ubio<-(arrayx>arrayy)-(arrayx<arrayy)

    xs<-X10[sam1]
    ys<-X20[sam2]
    xi<-ind10[sam1]
    yi<-ind20[sam2]
    matx<-matrix(rep(xs,n2),ncol=n2)
    matix<-matrix(rep(xi,n2),ncol=n2)
    maty<-matrix(rep(ys,n1),nrow=n1,byrow=T)
    matiy<-matrix(rep(yi,n1),nrow=n1,byrow=T)

    Usur<-(matix&matiy)*((matx>maty)-(matx<maty))-
    (matix&(!matiy))*(matx<=maty)+((!matix)&matiy)*(matx>=maty)

    UB[J+1,1]<-sum(Ubio[,,1])/(n1*n2)
    UB[J+1,2]<-sum(Ubio[,,2])/(n1*n2)
    UB[J+1,3]<-sum(Ubio[,,3])/(n1*n2)
    UB[J+1,4]<-sum(Usur)/(n1*n2)
    U[J+1]<-(sum(Ubio)+sum(Usur))/((K+1)*n1*n2)
  }
  ci1<-quantile(UB[,1], probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)
  ci2<-quantile(UB[,2], probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)
  ci3<-quantile(UB[,3], probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)
  ci4<-quantile(UB[,4], probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)
  rej<-(ci1[1]>0|ci1[2]<0)|(ci2[1]>0|ci2[2]<0)|(ci3[1]>0|ci3[2]<0)|
  (ci4[1]>0|ci4[2]<0)

  return(c(mean(U),sd(U),quantile(U, probs = c(0.025,0.975),na.rm=T),rej))

}
```

A.3 Bootstrap Function for Approach 2

```
Bootstrap2<-function(data10,data20,X10,X20,ind10,ind20,n1,n2,N,K){

  U<-rep(0,500)

  for (J in 0:499)
  {
    num1<-c(1:n1)
    num2<-c(1:n2)
    sam1<-sample(num1, size=n1, replace = TRUE, prob = NULL)
    sam2<-sample(num2, size=n2, replace = TRUE, prob = NULL)

    x<-data10[sam1,]
    y<-data20[sam2,]
    array1<-array(rep(x,n2),dim=c(n1,K,n2))
    arrayx<-aperm(array1,c(1,3,2))
    array2<-array(rep(y,n1),dim=c(n2,K,n1))
    arrayy<-aperm(array2,c(3,1,2))

    Ubio<-(arrayx>arrayy)-(arrayx<arrayy)

    xs<-X10[sam1]
    ys<-X20[sam2]
    xi<-ind10[sam1]
    yi<-ind20[sam2]
    matx<-matrix(rep(xs,n2),ncol=n2)
    matix<-matrix(rep(xi,n2),ncol=n2)
    maty<-matrix(rep(ys,n1),nrow=n1,byrow=T)
    matiy<-matrix(rep(yi,n1),nrow=n1,byrow=T)

    indc1<-1-xi
    surv1<-Surv(time=xs, event=indc1)
    km1<-summary(survfit(surv1~1))
    time1<-unlist(km1[2])
    survival1<-unlist(km1[6])
    len1<-length(time1)

    Ghat1 <- function(t){
      Ghat<-1
      if (len1>1)
      {
```

```
        for (l in 1:(len1-1))
        {
          Ghat<-Ghat-(1-survival1[l])*((t>time1[l])&(t<=time1[l+1]))
        }
      }
      Ghat<-Ghat-(1-survival1[len1])*(t>time1[len1])
      return(Ghat)
    }

    indc2<-1-yi
    surv2<-Surv(time=ys, event=indc2)
    km2<-summary(survfit(surv2~1))
    time2<-unlist(km2[2])
    survival2<-unlist(km2[6])
    len2<-length(time2)

    Ghat2 <- function(t){
      Ghat<-1
      if (len2>1)
      {
        for (l in 1:(len2-1))
        {
          Ghat<-Ghat-(1-survival2[l])*((t>time2[l])&(t<=time2[l+1]))
        }
      }
      Ghat<-Ghat-(1-survival2[len2])*(t>time2[len2])
      return(Ghat)
    }

    Usur<-matiy*pmax((matx>maty)/(Ghat1(maty)*Ghat2(maty)),0,na.rm=T)-
    matix*pmax((matx<maty)/(Ghat1(matx)*Ghat2(matx)),0,na.rm=T)

    U[J+1]<-(sum(Ubio)+sum(Usur))/((K+1)*n1*n2)
  }

  return(c(mean(U,na.rm=T),sd(U,na.rm=T),
  quantile(U, probs = c(0.025,0.975),na.rm=T)))
}
```

A.4 Bonferroni Test for Biomarkers Function

```
Bonferroni<-function(data10,data20,n1,n2,N,K)
{
  c<-rbind(data10,data20)
  r<-apply(c,2,order)
  ra<-r[(1:n1),]
  rsum<-apply(ra,2,sum)
  E<-n1*(N+1)/2
  SD<-sqrt(n1*n2*(N+1)/12)
  TS<-(rsum-E)/SD
  p<-pnorm(TS)
  rej<-as.numeric((p<0.025/(K+1))|(p>1-0.025/(K+1)))
  return(rej)
}
```

A.5 Bootstrap Function for Approach 2 (with Bonferroni Test)

```
Bootstrap2<-function(data10,data20,X10,X20,ind10,ind20,n1,n2,N,K){

  U<-rep(0,500)
  UB<-matrix(NA,nrow=500,ncol=4)

  for (J in 0:499)
  {
    num1<-c(1:n1)
    num2<-c(1:n2)
    sam1<-sample(num1, size=n1, replace = TRUE, prob = NULL)
    sam2<-sample(num2, size=n2, replace = TRUE, prob = NULL)

    x<-data10[sam1,]
    y<-data20[sam2,]
```

```
array1<-array(rep(x,n2),dim=c(n1,K,n2))
arrayx<-aperm(array1,c(1,3,2))
array2<-array(rep(y,n1),dim=c(n2,K,n1))
arrayy<-aperm(array2,c(3,1,2))

Ubio<-(arrayx>arrayy)-(arrayx<arrayy)

xs<-X10[sam1]
ys<-X20[sam2]
xi<-ind10[sam1]
yi<-ind20[sam2]
matx<-matrix(rep(xs,n2),ncol=n2)
matix<-matrix(rep(xi,n2),ncol=n2)
maty<-matrix(rep(ys,n1),nrow=n1,byrow=T)
matiy<-matrix(rep(yi,n1),nrow=n1,byrow=T)

indc1<-1-xi
surv1<-Surv(time=xs, event=indc1)
km1<-summary(survfit(surv1~1))
time1<-unlist(km1[2])
survival1<-unlist(km1[6])
len1<-length(time1)

Ghat1 <- function(t){
  Ghat<-1
  if (len1>1)
  {
    for (l in 1:(len1-1))
    {
      Ghat<-Ghat-(1-survival1[l])*((t>time1[l])&(t<=time1[l+1]))
    }
  }
  Ghat<-Ghat-(1-survival1[len1])*(t>time1[len1])
  return(Ghat)
}

indc2<-1-yi
surv2<-Surv(time=ys, event=indc2)
km2<-summary(survfit(surv2~1))
time2<-unlist(km2[2])
survival2<-unlist(km2[6])
len2<-length(time2)
```

```
    Ghat2 <- function(t){
      Ghat<-1
      if (len2>1)
      {
        for (l in 1:(len2-1))
        {
          Ghat<-Ghat-(1-survival2[l])*((t>time2[l])&(t<=time2[l+1]))
        }
      }
      Ghat<-Ghat-(1-survival2[len2])*(t>time2[len2])
      return(Ghat)
    }

    Usur<-matiy*pmax((matx>maty)/(Ghat1(maty)*Ghat2(maty)),0,na.rm=T)-
    matix*pmax((matx<maty)/(Ghat1(matx)*Ghat2(matx)),0,na.rm=T)

    UB[J+1,1]<-sum(Ubio[,,1])/(n1*n2)
    UB[J+1,2]<-sum(Ubio[,,2])/(n1*n2)
    UB[J+1,3]<-sum(Ubio[,,3])/(n1*n2)
    UB[J+1,4]<-sum(Usur)/(n1*n2)
    U[J+1]<-(sum(Ubio)+sum(Usur))/((K+1)*n1*n2)
  }

  ci1<-quantile(UB[,1], probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)
  ci2<-quantile(UB[,2], probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)
  ci3<-quantile(UB[,3], probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)
  ci4<-quantile(UB[,4], probs = c(0.025/(K+1),1-0.025/(K+1)),na.rm=T)
  rej<-(ci1[1]>0|ci1[2]<0)|(ci2[1]>0|ci2[2]<0)|(ci3[1]>0|ci3[2]<0)|
  (ci4[1]>0|ci4[2]<0)

  return(c(mean(U,na.rm=T),sd(U,na.rm=T),
  quantile(U, probs = c(0.025,0.975),na.rm=T),rej))
}
```

A.6 Data generating and test evaluation

```
Here, we take case 2 under alternative hypothesis setting as an example.
set.seed(80)
library("MASS")
library("survival")
library("stats")
n1<-100
n2<-100
N<-n1+n2
K<-3

#########
BGehan<-matrix(NA,nrow=500,ncol=6,,dimnames<-
list(c(1:500),c("mean","sd","95low","95up","Bon95low","Bon95up")))
Bkm<-matrix(NA,nrow=500,ncol=4,,dimnames<-
list(c(1:500),c("mean","sd","95low","95up")))
Bon<-matrix(NA,nrow=500,ncol=K,,dimnames<-
list(c(1:500),c("marker1","marker2","marker3")))
sigma3<-matrix(rep(0,K*K),nrow=K)
for (i in 1:K)
{
  sigma3[i,i]<-3
  sigma3[i,i-1]<-1
  sigma3[i-1,i]<-1
}

for (LOOP in 1:500)
{
#data
data10<-mvrnorm(n = n1, mu=c(0,1,2), Sigma=sigma3)
T10<-rexp(n1,0.5)
T10<-T10+0.03*data10[,2]
C10<-runif(n1,0,3)
X10<-pmin(T10,C10)
ind10<-(T10>=C10)

data20<-mvrnorm(n = n2, mu=c(-1,3,2.5), Sigma=sigma3)
T20<-rexp(n2,0.4)
T20<-T20+0.03*data20[,2]
C20<-runif(n2,0,3)
X20<-pmin(T20,C20)
ind20<-(T20>=C20)
```

```
#bootstrap
BGehan[LOOP,]<-Bootstrap1(data10,data20,X10,X20,ind10,ind20,n1,n2,N,K)
Bkm[LOOP,]<-Bootstrap2(data10,data20,X10,X20,ind10,ind20,n1,n2,N,K)
#Bonferroni
Bon[LOOP,]<-Bonferroni(data10,data20,n1,n2,N,K)
}

#################


Bon0<-BGehan[,5:6]
Bon1<-(Bon0[,1]>0)|(Bon0[,2]<0)

mean(BGehan[,1])
mean(Bkm[,1])
mean(BGehan[,2])
mean(Bkm[,2])
mean((BGehan[,3]>0)|(BGehan[,4]<0))
mean((Bkm[,3]>0)|(Bkm[,4]<0))

mean(Bon[,1]|Bon[,2]|Bon[,3]|Bon1)
```

Bibliography

[1] T. R. Fleming, and D. L. DeMets. (1996). Surrogate end points in clinical trials: Are we being misled? *Ann Intern Med* **125**, 605-613.

[2] P. Huang, R. F. Woolson and P. C. OBrien. (2008). A rank-based sample size method for multiple outcomes in clinical trials, *Statistics in Medicine* **27**, 3084-3104.

[3] R. A. Fisher. (1996). The fiducial argument in statistical inference, *Annals of Eugenics* **6**, 391398.

[4] E. Brunner, U. Munzel, and M. L. Puri. (2002). The multivariate nonparametric BehrensFisher problem, *Journal of Statistical Planning and Inference* **108**, 37-53.

[5] P. C. O'Brien, N. L. Geller. (1997). Interpreting tests for efficacy in clinical trials with multiple endpoints, *Controlled Clinical Trials* **18**, 591-602222-227.

[6] L. Acion, J. J. Peterson, S. Temple, S. Arndt. (2006). Probabilistic index: an intu-

itive non-parametric approach to measuring the size of treatment effects, *Statistics in Medicine* **25**, 591-602.

[7] P. Huang, B. C. Tilley, R. F. Woolson, and S. Lipsitz. (2005). Adjusting O'Brien's test to control Type I error for the generalized nonparametric Behrens-Fisher problem. *Biometrics* **61**, 532-539.

[8] P. C. O'Brien. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079-1087.

[9] S. C. Cheng, L. J. Wei, and Z. Ying. (1995). Analysis of Transformation models with censored data, *Biometrika* **82**, 835-845

[10] E. L. Kaplan, and P. Meier. (1958). Nonparametric estimation from incomplete observations, *JASA* **53**, 457-481

Vita

Shuo Xu received the B.S. degree in mathematics and applied mathematics from Nanjing University in 2012, enrolled in the Biostatistics Sc.M. program at Johns Hopkins University in 2012, and has been an exchange student in National Tsing-Hua University (Taiwan) in 2010. She won the Nanjing University Honor Student in 2009, received a Nanjing University Third Class People's Scholarship in 2009 and Johns Hopkins Bloomberg School of Public Health 75% Master's Tuition Scholarship in 2013, and was an Outstanding Member of Nanjing University Red-Cross and an approved first-aid provider for Jiangsu Red-Cross. Her current research focuses on survival Analysis and surrogate endpoints, her undergraduate research focused on congruence theory in number theory, and her co-authored paper with Z. Zhang "*Cryptanalysis and Improvement of a Concurrent Signature, Scheme Based on Identity*" was published on The 2nd IEEE International Conference on Software Engineering and Service Sciences (2011 Beijing, China).

VITA